A Framework for Interestingness Measures for Association Rules with Discrete and Continuous Attributes Based on Statistical Validity

Izwan Nizal Mohd Shaharanee(✉) and Jastini Mohd Jamil

School of Quantitative Sciences, Universiti Utara Malaysia, UUM, 06010 Sintok, Malaysia
{nizal,jastini}@uum.edu.my

Abstract. Assessing rules with interestingness measures is the pillar of successful application of association rules discovery. However, association rules discovered are large in number, some of which are not considered as interesting or significant for the application at hand. In this paper, we present a systematic approach to ascertain the discovered rules, and provide a precise statistical approach supporting this framework. Furthermore, considering that many interestingness measures exist, we propose and compare two established approaches in selecting relevant attributes for the rules prior to rule generation. The proposed strategy combines data mining and statistical measurement techniques, including redundancy analysis, sampling and multivariate statistical analysis, to discard the non-significant rules. In addition to that, we consider real world datasets which are characterized by the uniform and non-uniform data/items distribution with mixture of measurement level throughout the data/items. The proposed unified framework is applied on these datasets to demonstrate its effectiveness in discarding many of the redundant or non-significant rules, while still preserving the high accuracy of the rule set as a whole.

Keywords: Data mining · Interesting rules · Statistical analysis

1 Introduction

Data mining or knowledge discovery from data (KDD) is known for its capabilities in offering systematic ways in acquiring useful rules and patterns from large quantities of data. The rules derived from data mining application are considered interesting and useful if they are comprehensible, valid on tests and new data with some degree of certainty, potentially useful, actionable, and novel [1]. [2] claims that the majority of data mining/machine learning type patterns are rule based in nature with a well defined structure, such as rules derived from decision trees and association rules. The most common patterns that can be evaluated by interestingness measures include association rules, classification rules, and summaries [3]. Association rule mining is one of the most popular data mining techniques widely used for discovering interesting associations and correlations between data elements in a diverse range of applications [4]. The association rule mining techniques are different from each other, but a commonality that remains is that all the frequent patterns are first extracted and then

T. Dillon (Ed.): IFIP AI 2015, IFIP AICT 465, pp. 119–128, 2015.
DOI: 10.1007/978-3-319-25261-2_11

association rules are formed from such patterns. Frequent pattern extraction plays an important part in generating good and interesting rules, and is considered the most difficult and complex task. Different methods have been proposed for discovering interesting rules from data and have been categorized into three main classes, namely objective, subjective and semantic measures [1-3].

Our work in the area of rules interestingness measures is motivated by the objective interestingness measures which are based on probability theory, statistics and information theory. Various objective interestingness criteria have been used to limit the nature of rules extracted, as explained in [3, 6]. Works such as [7, 8] have proposed and successfully developed two approaches, namely multiple support and relative support for generating rules for significant rare data that appears infrequent in the database but is highly associated with specific data. Mutual Information and J-Measures are common information theory approaches in objective interestingness measure [9]. A number of researchers have anticipated an assessment on pattern discovery by applying a statistical significance test as discussed in [6].

Assessing whether a rule satisfies a particular constraint is accompanied by a risk that the rule will satisfy the constraint with respect to the sample data but not with respect to the whole data distribution [10]. As such, the rules may not reflect the "real" association between the underlying attributes. The hypotheses reflected in the generated rules must be validated by a statistical methodology for them to be useful in practice, because the nature of data mining techniques is data driven [11]. However, even if the rules satisfy appropriate statistical tests, it can still be the case that the underlying association is caused purely by a statistical coincidence [12].

The contributions of the work presented in this paper, is in developing systematic ways to verify the usefulness of rules obtained from association rules mining using statistical analysis. A unified framework is proposed, that combines several techniques to access the quality of rules, and remove any redundant and unnecessary rules. Initial ideas and preliminary results were presented earlier in [6, 13]. Several extensions and refinements took place in regards to the method being applicable to more realistic datasets including complex data types, infrequent items and uneven attribute value distribution. Furthermore, a comparison of the statistical measure used in our framework with the popular Mutual Information measure is included. The rest of the paper is organized as follows. Section 2, briefly overviews some related works and defines the problem of ascertaining the discovered rules. In Section 3, we describe our proposed framework. The framework is evaluated using real world datasets and some experimental findings and explanation are given in Section 4. Section 5 concludes the paper and describes our ongoing works in this field of study.

2 Related Works

Association rule mining in its most fundamental structure is to discover interesting relationships among items in a given dataset under minimum support and confidence conditions. Commonly used example is in market basket analysis, where an association rule $X \Rightarrow Y$ means if a consumer buys the set of items X, then he/she probably also buys items Y. These items are typically referred to as itemsets [14]. The problem of finding association rules $X \Rightarrow Y$ was first introduced in [5, 15] as a data mining task of finding frequently co-occurring items in a large Boolean transaction database. Let $I = \{i_1, i_2, ..., i_m\}$ be a set of items. Each transaction T is a set of items, such that $T \subseteq I$. An association rule is a condition of the form of $X \Rightarrow Y$ where $X \subseteq I$ and $Y \subseteq I$ are two sets of items. The support of a rule $X \Rightarrow Y$ is the number of transactions that contain both X and Y, while the confidence of a rule $X \Rightarrow Y$ is the number of transactions containing X, that also contain Y.

In this research we employed an efficient breadth-first method in generating candidate set called Apriori algorithm [16]. The generation of all possible rules was essential in ascertaining the quality of the rules. As this established approach is based on the user specifying the constraints such as support and confidence that must be satisfied. Still, [7, 8] argue that for a real large database that is often comprised of either relatively frequent/infrequent items, using multiple and relative support should be considered.

The rules satisfying the standard support and confidence constraints are often too numerous to be utilized efficiently and effectively for the application at hand [17]. Many patterns from the frequent pattern set are often redundant. Thus we discussed in detailed two useful approaches in handling this redundant problem in [6]. In the datasets where there is a predefined class label (i.e. classification tasks), frequent pattern mining can contribute to discovering strong associations between occurring attribute and class values. In [18] the potential usage of frequent pattern mining for classification problem was investigated and successfully applied to the problem. Their approach discovered classification rules by directly discovering the frequent patterns from the datasets with predefined class labels. The results reported were promising since the discovered knowledge model had high accuracy and efficiency for the classification problem.

3 Proposed Method

Although there are various criteria in determining the usefulness of rules [1, 2, 17] the measures usually just reflect the usefulness of rules with respect to the specific database being observed [10]. The data mining approaches consider the whole search space to find all possible pattern/rules satisfying specific criteria (i.e. association rules). While these criteria, offer some constrains in discovering strong patterns/rules, many misleading, uninteresting and insignificant rules in that domains may still be produced [1]. This problem arises because some association rules are discovered due to pure coincidence resulting from certain randomness in the particular dataset being analyzed. Statistics has previously addressed the issues of how to separate out the random effects to determine if the measured association (or difference in other areas) is significant [22, 23]. Thus additional measures based on statistical independence and correlation analysis are needed to ensure that the results have a sound statistical basis and are not purely random coincidence. The statistical approach offers a firm way of identifying significant rules that are statistically valid. Therefore, the motivation behind our proposed method is to investigate how data mining and statistical measurement techniques can be combined to arrive at more reliable and interesting set of rules. Generally speaking we interpret interesting rules as those rules that have a sound statistical basis and are not redundant. Such an approach requires sampling process, hypothesis development, model building and finally a measurement using statistical analysis techniques to verify and ascertain the usefulness and quality of the rules discovered. This will filter out the redundant, misleading, random and coincidentally occurring rules, while at the same time the accuracy of the rule set will still be sustained.

3.1 Conceptual Framework

Figure 1 shows the proposed framework. The dataset is first divided into two partitions. The first partition is used for association rule generation and statistical evaluation, while the second partition acts as a sample data drawn from the database, used to verify the accuracy of discovered rules. To ensure clean and consistent data, standard preprocessing techniques are applied. These preprocessing techniques include the removal of missing values and discretization of attributes with continuous values. As the next step, we determine the relevance of attributes by classifying their importance to characterize an association. A powerful technique for this purpose is the Symmetrical Tau [19], which is a statistical-heuristic feature selection criterion. It measures the capability of an attribute in predicting the class of another attribute. The measure is based on the probabilities of one attribute value occurring together with the value of the second attribute. [19] define the Symmetrical Tau measure for the capability of input attribute in predicting the class attribute. Higher values of the Tau measure would indicate better discriminating criterions (features) for the class that is to be predicted in the domain. Symmetrical Tau has many more desirable properties in comparison to other feature selection techniques, as was reported in [19].

In Section 4.2 we evaluate the capabilities of Symmetrical Tau as the determinant of relevance attributes by comparing it with an information-theoretic measure. The information-theoretic measures are principally comprehensible and useful since they can be interpreted in terms of information. For a rule interestingness measure, the relation is interesting when the antecedent provides a great deal of information about the consequent [20]. Although several information-theoretic measures exist, we only compared Symmetrical Tau with Mutual Information measurement technique. The Mutual Information is based on information theory to evaluate rules. This approach describes how much information one random variable tells about another one [21]. The definition of Mutual Information is based on [9]. The features selection technique is utilized in our approach to provide the relative usefulness of attributes in predicting the value of the class attribute, and discard any of the attributes whose relevance value is low. This would prevent the generation of rules which then would need to be discarded anyway once it was found that they comprise of some irrelevant attributes.

The rules are then generated based on the minimum support and confidence framework. However, the application of minimum support assumes that all items in the data are of the same nature and/have similar frequency in the database. This will encounter problems given that the frequency distribution of the attribute values or items in the dataset can be significantly different [7]. [8] assert that the data distribution in database may somehow occur either relatively frequently or not, uniformly or non-uniformly distributed according to the characteristics of the database. In response to this rare items problem, several researchers proposed and successfully developed a solution such as multiple minimum support [7] and relative support [8]. For comparison purpose, we apply both relative support and classical Apriori algorithm framework for association rules mining generation. This is to ensure that we treat the dataset that contains rare items correctly, and this will be demonstrated in the experiments provided in Section 4.3.

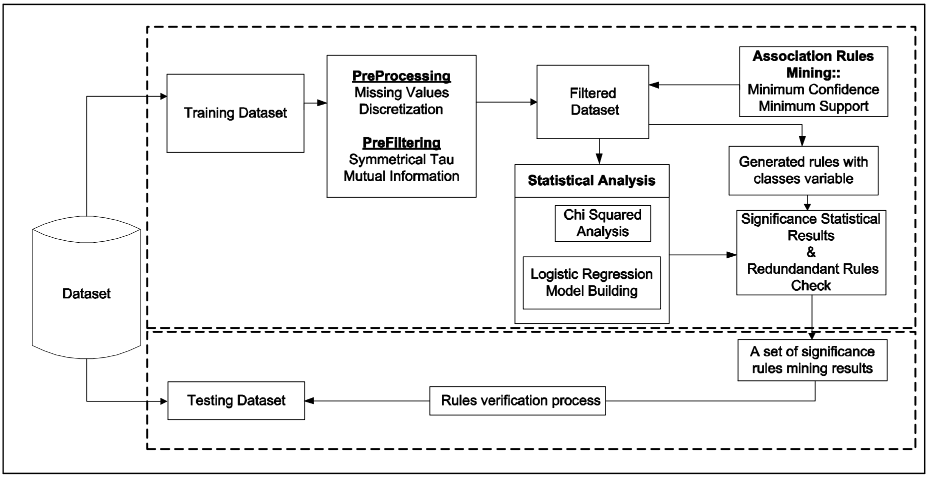

Fig. 1. Proposed framework for rule interestingness analysis.

The discovered rules are then ascertained with statistical techniques. For associations between categorical and continuous variables there are several inferential methods involved. Chi-squared analysis is often used to measure the correlation between items. For a given chi-squared values it can be used to determine if the correlation is statistically significant [1]. The logistic regression methods have become an integral component of any data analysis concerned with describing the relationship between a target variable and one or more input variables [22]. Logistics regression is used to estimate the probability that a particular outcome will occur. The coefficients are estimated using a statistical technique called maximum likelihood estimation. These coefficient values are useful in testing the statistical significance of input variables towards target variables [23]. The interpretation of regression coefficient in terms of odd ratios is a familiar concept in analysis of categorical data [23]. The selection of logistic regression model involves two competing goals: the model should be complex enough to fit the data well, while at the same time simpler models are preferred since they are easier to interpret and are expected to have better generalization [22].

We also use some constraint measurement techniques in order to discard the existence of redundant rules [6]. The combination of these rule ascertaining strategies will facilitate the association rule mining framework to determine the right and high quality rules.

4 Experimental Results

The evaluation of the unification framework is performed using the Adult, Iris and Wine dataset, which are real world datasets obtained form UCI Machine Learning Repository. Since all the datasets used are supervised which reflects a classification problem, we have chosen the target variable as the right hand side/consequence of the association rules discovered during association rule mining analysis. In this section, we first show how we handle continuous and discrete attributes. Then we compare two established features selection algorithms namely Symmetrical Tau and Mutual Information in term of their feature subset selection capabilities. Next, we discuss the effect of the frequent and infrequent item in dataset towards the framework. Finally we demonstrate the whole performance of the framework towards both datasets.

4.1 Discrete and Continuous Attributes

For many real-world problems, the forms of the input and target attributes emerge form wide range of measurement levels. In handling these types of attributes, we apply the binning approach in improving the boundary of the continuous variables. These bound are created to reflect the upper and lower values for the input variables [24]. For all continuous attributes in Adult, Iris and Wine, we apply equal depth binning approach methods. This equal depth binning approach will ensure that, we have a manageable data sizes by reducing the number of distinct values per attributes [1]. Other discrete attributes in Adult dataset were preserved in original state.

4.2 Comparing Symmetrical Tau (ST) with Mutual Information (MI)

ST and MI are capable of defining irrelevant attributes; they are different from each other in terms of their approach as aforementioned in Section 3.1. Throughout the

Table 1. Comparison between ST and MI for Adult Dataset (Initial Proportion)

# of Values	Variables	ST Values		# of Values	Variable	MI Values
7	Marital Status	0.1448		6	Relationships	0.1662
6	Relationship	0.1206		7	Marital Status	0.1575
6	Capital Gain	0.0706		16	Education	0.0934
8	Education Number	0.0688		14	Occupation	0.0932
16	Education	0.0528		8	Education Number	0.0900
2	Sex	0.0470		10	Age	0.0894
14	Occupation	0.0469		10	Hours Per Week	0.0545
10	Age	0.0432		6	Capital Gain	0.0475
5	Capital Loss	0.0361		2	Sex	0.0374
10	Hours Per Week	0.0354		5	Capital Loss	0.0238
7	Work Class	0.0166		7	Work Class	0.0171
5	Race	0.0085		41	Native Country	0.0093
41	Native Country	0.0077		5	Race	0.0083
10	FNLWGT	0.0002		10	FNLWGT	0.0002

experiment as shown in Table 1, we found that MI approach favors variables with more values. This observation is in accord with [20]. On the contrary, the procedure based on ST produces a more stable variables selection which is not in favor to any specific variables criterion. This is in agreement with the claim in [19, 25, 26], of ST being fair towards handling of multi-valued variables.

4.3 Unified Target Data and Rare Target Data Problems

As mention in Section 3.1, [8] assert that the data distribution in database may somehow occur either relatively frequently or not according to the database's characteristics. In response to this, we compared the dataset that contains both unified target data and rare target data.

Table 2. Rules accuracy for Adult data

Experimental Approaches	Dataset Description	Rule #	Type of Statistical Analysis	Accuracy	
				Training	Testing
Initial Proportion	Training : 30162 records Testing : 15060 records	164	Initial rules	86.75%	86.87%
		53	Statistical Analysis	87.73%	87.92%
		42	Redundancy Check	87.99%	88.13%
Balanced Data	Training : 30162~15016 records Testing : 15060 records	421	Initial rules	71.55%	60.56%
		51	Statistical Analysis	73.87%	58.28%
		30	Redundancy Check	74.00%	63.80%
Replication Data	Training : 30162~45178 records Testing : 15060 records	255	Initial rules	71.65%	59.86%
		51	Statistical Analysis	73.64%	58.28%
		32	Redundancy Check	73.61%	61.70%
*Multiple Support	Training: 30162 records Testing : 15060 records	164	Initial rules	86.75%	86.87%
		53	Statistical Analysis	87.73%	87.92%
		42	Redundancy Check	87.99%	88.13%
	*5% as 2nd support	42+*4	Redundancy Check	87.34%	87.47%

Table 2 shows four experiments done for Adult dataset. For the Adult dataset, we have limited the consequence of the rules to be either Income > 50K or Income =<50K. The initial proportion of this target data is unbalanced, making the target data for Adult dataset consist of an infrequent target value (rare target data).

Firstly, we apply the minimum support approach based on the initial proportion of training and test data. Next, we show the results when we have balanced the training dataset, so that we can have a similar proportion between the training and testing data. Then, we made some replication of records in the training dataset. This replication process has generated additional records for training data so that any value from the set of target values has a more similar frequency of occurrence in the training dataset and this will represent a similar proportion between each target item. Finally, we generated the rules based on relative support approach proposed by [8]. Based on results obtained, we conclude that, for a rare target data, the most suitable approach in generating the rules are by applying the relative support. This agrees with [8], which have successfully applied the relative support in identifying the strong co-relation of significant rare data items compared to classical minimum support approach. While the work presented in [8] purposely aims for the efficiency of rare item rule generation, our proposed framework, demonstrated its capabilities in ascertaining the generated rules. Table 3 shows the rules obtained for Iris and Wine dataset as in these datasets contained balanced target values (unified target data).

Table 3. Rules accuracy for Iris and Wine data.

Dataset Name	Dataset Description	Rule #	Type of Statistical Analysis	Accuracy	
				Training	Testing
Iris	Training : 90 records Testing : 60 records	52 22	Initial Rules Redundancy Check	92.86% 88.15%	90.99% 85.29%
Wine	Training : 107 records Testing : 71 records	195 17 16	Initial Rules Statistical Analysis Redundancy Check	87.53% 85.07% 85.07%	79.44% 81.98% 81.98%

4.4 Overall Framework Performance

Taking in the whole dataset as input would produce a large number of rules, many of which are caused by the presence of irrelevant attributes. Since the ST has more advantageous assets in comparison to MI, ST feature selection criterion is used earlier in the process to remove any irrelevant attributes. This would prevent the generation of rules that comprise of some irrelevant attributes. Hence in this experiment it is not necessary to use ST to further verify the rules as the rules were created from the attribute subset considered as relevant by the measure, as was done in [6, 13]. The attributes were ranked according to their decreasing ST and a relevance cut-off point was picked. In this experiment, the cut off value was picked based on the significant difference between the ST values in decreasing order. The significant difference was considered to occur in the ranking at the position where that attribute's ST value is less than half of the previous attribute's ST value in the ranking. At this point and below in the ranking, all attributes are considered as irrelevant, as is indicated in Table 1. For example, for Adult dataset, the relevance cutoff value is 0.0166. This is due to the ST value of attribute 'Hours Per Week' being more than double of the ST value for attribute 'Work Class'. Thus, the subset of data consists now of 10 attributes: Marital Status, Relationship, Capital Gain, Education Number, Education, Sex , Occupation, Age, Capital Loss and Hours Per Week. We proceed with the application of an association rule mining algorithm and verification of the extracted rules through statistical analysis. As discussed in 4.3, we concluded that relative support approach is capable of generating rules form a rare target data as in the Adult dataset. On the right hand side of Table 2, we show the progressive difference in the number of rules generated as statistical analysis and redundancy checks are being utilized. The combination of statistical significance analysis and redundant analysis provided proper ways in discarding non-significant rules, which is a significant reduction in the overall complexity of the rule set. From Table 2 we can also see that this great reduction of rules was not at a cost of a significant reduction in accuracy.

In Table 2, based on the statistical and redundant rules analysis performed at rules obtained from relative support approach, we managed to get 42+4 rules as our final significance rules. Input variables namely Marital Status, Relationship, Education, Sex, Occupation, Age, Capital Loss and Hours per Week are selected for the final rules. As depicted in Table 3, the result for Iris and Wine dataset also show no significant deterioration in accuracy with the reduced rule set. As to gauge the effect of rules accuracy on difference set of partitioning for each dataset, k-fold cross validation approach has being utilized, to ensure that we obtained relatively low bias and variance [1]. Based on the experimental results, we have found that the average reduction in the accuracy of the rules set is minor in comparison to the major reduction in the complexity of the rule set.

5 Conclusions and Future Works

This paper has presented a framework to ascertain the quality of the rules discovered from association rule mining which has a huge amount of rules and complex attributes measurement levels with an integrated statistical and heuristic measurement technique. The experimental results show that, this framework managed to reduce a large number of non-significant and redundant rules while at the same time relatively high accuracy was preserved. This indicates the potential of the framework in providing significant rules when applied to the structured or relational data. As part of our ongoing works, we intend to use the proposed framework to ascertain more complex rules which are discovered from semi-structured data.

References

1. Han, J., Kamber, M.: Data mining : concepts and techniques. Morgan Kaufmann Publishers, San Francisco (2001)
2. McGarry, K.: A survey of interestingness measures for knowledge discovery. Knowl. Eng. Rev. **20**, 39–61 (2005)
3. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: A survey. ACM Comput. Surv. **38**, 9 (2006)
4. Zhang, H., Padmanabhan, B., Tuzhilin, A.: On the discovery of significant statistical quantitative rules. In: Proceedings of the 10th ACM SIGKDD International Conference On Knowledge Discovery And Data Mining. ACM, New York (2004)
5. Agrawal, R., Imieliski, T., Swami, A.: Mining association rules between sets of items in large databases. In: SIGMOD Rec., vol. 22, pp. 207–216 (1993)
6. Shaharanee, I.N.M., Hadzic, F., Dillon, T.S.: Interestingness of association rules using symmetrical tau and logistic regression. In: Nicholson, A., Li, X. (eds.) AI 2009. LNCS, vol. 5866, pp. 422–431. Springer, Heidelberg (2009)
7. Bing, L., Wynne, H., Yiming, M.: Mining association rules with multiple minimum supports. In: Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, California (1999)
8. Yun, H., Ha, D., Hwang, B.: Ho Ryu, K.: Mining association rules on significant rare data using relative support. Journal of Systems and Software **67**, 181–191 (2003)

9. Tan, P.N., Kumar, V., Srivastava, J.: Selecting the right interestingness measure for association patterns. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Alberta (2002)
10. Webb, G.I.: Discovering Significant Patterns. Machine Learning, 1–33 (2007)
11. Goodman, A., Kamath, C., Kumar, V.: Data Analysis in the 21st Century. Stat. Anal. Data Min. **1**, 1–3 (2008)
12. Aumann, Y., Lindell, Y.: A Statistical Theory for Quantitative Association Rules. J. Intell. Inf. Syst. **20**, 255–283 (2003)
13. Shaharanee, I.N.M., Dillon, T.S., Hadzic, F.: Ascertaining association rules using statistical analysis. In: Proceeding of the 2009 International Symposium on Computing, Communication and Control, Singapore (2009)
14. Philippe, L., Patrick, M., Benoît, V., Stéphane, L.: On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. European Journal of Operational Research **184**, 610–626 (2008)
15. Aggarwal, C.C., Yu, P.S.: A new framework for itemset generation. In: Book a new framework for itemset generation. Series A new framework for itemset generation. ACM, New York (1998)
16. Toivonen, H.: Sampling large databases for association rules. In: Proceedings of the 22th International Conference on Very Large Data Bases. Morgan Kaufmann Publishers Inc. (1996)
17. Lavrač, N., Flach, P.A., Zupan, B.: Rule evaluation measures: a unifying view. In: Džeroski, S., Flach, P.A. (eds.) ILP 1999. LNCS (LNAI), vol. 1634, pp. 174–185. Springer, Heidelberg (1999)
18. Cheng, H., Yan, X., Han, J., S., Y.P.: Direct discriminative pattern mining for effective classification. In: Proceedings of the 24th International Conference on Data Engineering, ICDE 2008, pp. 169–178 (2008)
19. Zhou, X.J., Dillon, T.S.: A statistical-heuristic feature selection criterion for decision tree induction. IEEE Transaction on Pattern Analysis and Machine Intelligence **13** (1991)
20. Julien, B., Fabrice, G., Regis, G., Henri, B.: Using information-theoretic measures to assess association rule interestingness. In: Proceedings of the 5th IEEE International Conference on Data Mining. IEEE Computer Society (2005)
21. Lotfi, S., Sadreddini, M.H.: Mining fuzzy association rules using mutual information. In: International MultiConference of Engineers and Computer Scientists, vol. 1, Hong Kong (2009)
22. Agresti, A.: An Intro to Categorical Data Analysis. Wiley-Interscience, New York (2007)
23. Hosmer, D.W., Lemeshow, S.: Applied logistic regression. Wiley, New York (1989)
24. Dillon, T.S., Hossain, T., Bloomer, W., Witten, M.: Improvements in supervised BRAINNE: a method for symbolic data mining using neural networks. In: Seventh Conference on Database Semantics, vol. 124, pp. 67–88. Chapman & Hall, Switzerland (1998)
25. Shaharanee, I., Hadzic, F.: Evaluation and optimization of frequent, closed and maximal association rule based classification. Stat. Comput. **23**, 1–23 (2013)
26. Shaharanee, I., Hadzic, F., Dillon, T.: Interestingness measures for association rules based on statistical validity. Knowl.-Based Syst. **24**(3), 386–392 (2011)